BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Chromatin Regulator Genes

Cross-Reference to Related Applications

This application is a continuation-in-part of U.S. Application No. 08/945,988, filed November 10, 1997, which is the national phase entry application of PCT/EP96/01818, filed May 2, 1996, claiming priority to German Application No. DE 195 16 776.7, filed May 10, 1995. These applications are incorporated herewith by reference in their entirety.

Background of the Invention

Field of the Invention

The present invention relates to genes which play a part in the structural and functional regulation of chromatin, and their use in therapy and diagnosis.

Related Art

Higher-order chromatin is essential for epigenetic gene control and for the functional organization of chromosomes. Differences in higher-order chromatin structure have been linked with distinct covalent modifications of histone tails which regulate transcriptional 'on' or 'off' states and influence chromosome condensation and segregation .

Histones constitute a highly conserved family of proteins (H3, H4, H2A, H2B, H1) which are the major components of eucaryotic chromatin structure. Histones compact genomic DNA into basic repeating structural units, the nucleosomes. In addition to their DNA packaging function, histones have been proven to be integral components of the molecular machinery that regulates gene expression.

Post-translational modifications of histone N-termini, particularly of H4 and H3, are well-documented and have functionally been characterized as changes

in acetylation, phosphorylation and, most recently, methylation. In contrast to the large number of described histone acetyltransferases (HATs) and histone deacetylases (HDACs), genes encoding enzymatic activities that regulate phosphorylation or methylation of histone N-termini are only beginning to be identified. Moreover, the interdependence of the different histone tail modifications for the integration of transcriptional output or higher-order chromatin organization is currently not understood.

Overall, there is increasing evidence that the regulation of normal and aberrant cellular proliferation is not only affected on the transcriptional level, but that also a higher level of regulation is involved, i.e., the organization of chromatin structure through the modification of histone molecules. The determination of the proteins and the molecular mechanisms involved in histone modification will contribute to the understanding of the cellular proliferation program and will thus shed led light on the mechanisms involved in aberrant proliferation occurring in tumor formation and progression .

The functional organization of eucaryotic chromosomes in centromeres, telomeres and eu- and heterochromatic regions is a crucial mechanism for ensuring exact replication and distribution of genetic information on each cell division. By contrast, tumor cells are frequently characterized by chromosomal rearrangements, translocations and aneuploidy (Solomon, *et al., Science 254*:1153-1160 (1991); Pardue, *Cell 66*:427-431 (1991)).

Although the mechanisms which lead to increased chromosome instability in tumor cells have not yet been clarified, a number of experimental systems, beginning with telomeric positional effects in yeast (Renauld, *et al., Genes & Dev. 7*: 1133-1145 (1993); Buck and Shore, *Genes & Dev. 9*:370-384 (1995); Allshire, *et al., Cell 76*:157-169 (1994)), via positional effect variegation (PEV) in *Drosophila* (Reuter and Spierer, *BioEssays 14*:605-612 (1992)), and up to the analysis of translocation fracture points in human leukaemias (Solomon, *et al., Science 254*:1153-1160 (1991); Cleary, *et al., Cell 66*:619-622 (1991)), have

made it possible to identify chromosomal proteins which are involved in causing deregulated proliferation.

First, it was found that the overexpression of a shortened version of the SIR4-protein leads to a longer life in yeast (Kennedy, *et al., Cell 80*:485-496 (1995)). Since SIR proteins contribute to the formation of multimeric complexes at the stationary mating type loci and at the telomere, it could be that overexpressed SIR4 interferes with these heterochromatin-like complexes, finally resulting in uncontrolled proliferation. This assumption accords with the frequency of occurrence of a deregulated telomere length in most types of human cancer (Counter, *et al., Embo. J. 11*:1921-1928 (1992)).

Second, genetic analyses of PEV in *Drosophila* have identified a number of gene products which alter the structure of chromatin at heterochromatic positions and within the homeotic gene cluster (Reuter and Spierer, *BioEssays 14*:605-612 (1992)). Mutations of some of these genes, such as *modulo* (Garzino, *et al., Embo J. 11*:4471-4479 (1992)) and *polyhomeotic* (Smouse and Perrimon, *Dev. Biol. 139*:169-185 (1990)), can cause deregulated cell proliferation or cell death in *Drosophila*.

Third, mammalian homologues of both activators, e.g., *trithorax* or *trx*-group, and also repressors, e.g., *polycomb* or *Pc*-group, of the chromatin structure of homeotic *Drosophila* selector genes have been described. Among these, human *HRX/ALL-1* (*trx*-group) has been shown to be involved in leukaemogenesis induced by translocation (Tkachuk, *et al., Cell 71*:691-700 (1992); Gu, *et al., Cell 71*:701-708 (1992)), and it has been shown that the overexpression of murine *bmi* (*Pc*-group) leads to the formation of lymphomas (Haupt, *et al., Cell 65*:753-763 (1991); Brunk, *et al., Nature 353*:351-355 (1991); Alkema, *et al., Nature 374*:724-727 (1995)). A model for the function of chromosomal proteins leads one to conclude that they form multimeric complexes which determine the degree of condensation of the surrounding chromatin region depending on the balance between activators and repressors in the complex (Locke, *et al., Genetics 120*:181-198 (1988)). A shift in this equilibrium, caused by overexpression of one

of the components of the complex, exhibited a new distribution of eu- and heterochromatic regions (Buck and Shore, *Genes & Dev. 9*:370-384 (1995); Reuter and Spierer, *BioEssays 14*:605-612 (1992); Eissenberg, *et al., Genetics 131*:345-352 (1992)) which can destabilize the chromatin structure at predetermined loci, and lead to a transition from the normal to the transformed state.

In spite of the characterization of *HRX/ALL-1* and *bmi* as protooncogenes which are capable of changing the chromatin structure, knowledge of mammalian gene products which interact with chromatin is still very limited. By contrast, by genetic analyses of PEV in *Drosophila*, about 120 alleles for chromatin regulators have been described (Reuter and Spierer, *BioEssays 14*:605-612 (1992)).

Recently, a carboxy-terminal region was identified with similarity in the sequence to a positive (*trx* (*trx*-group)) and a negative (*E(z)* (*Pc*-group)) *Drosophila* chromatin regulator (Jones and Gelbart, *MCB 13(10)*:6357-6366 (1993)). Moreover, this carboxy terminus is conserved in *Su(var)*3-9, a member of the *Su(var)* group, and a dominant suppressor of chromatin distribution in *Drosophila* (Tschiersch, *et al., Embo J. 13(16)*:3822-3831 (1994)).

Genetic screens for suppressors of position effect variegation (PEV) in *Drosophila* and *S. pombe* have identified a subfamily of approximately 30-40 loci which are referred to as *Su(var)*-group genes. Interestingly, several histone deacetylases, protein phosphatase type 1 and S-adenosyl methionine synthetase have been classified as *Su(var)s*. In contrast, *Su(var)2-5* (which is allelic to *HP1*), *Su(var)3-7* and *Su(var)3-9* encode heterochromatin-associated proteins. *Su(var)* gene function thus suggests a model in which modifications at the nucleosomal level may initiate the formation of defined chromosomal subdomains that are then stabilized and propagated by heterochromatic SU(VAR) proteins. *Su(var)3-9* is dominant over most PEV modifier mutations, and mutants in the corresponding *S. pombe clr4* gene disrupt heterochromatin association of other modifying factors and result in chromosome segregation defects. Recently, human (*SUV39H1*) and murine (*Suv39h1* and *Suv39h2*) *Su(var)3-9* homologues have been isolated . It

has been shown that they encode heterochromatic proteins which associate with mammalian *HP1*. The SU(VAR)3-9 protein family combines two of the most evolutionarily conserved domains of 'chromatin regulators': the chromo and the SET domain. Whereas the 60 amino acid chromo domain represents an ancient histone-like fold that directs eu- or heterochromatic localizations, the molecular role of the 130 amino acid SET domain has remained enigmatic. Overexpression studies with human *SUV39H1* mutants indicated a dominant interference with higher-order chromatin organization that, surprisingly, suggested a functional relationship between the SET domain and the distribution of phosphorylated (at serine 10) H3.

The experiments of the present invention show that mammalian *SUV39H1* or *Suv39h* proteins are SET domain-dependent, H3-specific histone methyltransferases (HMTases) which selectively methylate lysine 9 of the H3 N-terminus. Methylation of lysine 9 negatively regulates phosphorylation of serine 10 and reveals a 'histone code' that appears intrinsically linked to the organization of higher-order chromatin.

Summary of the Invention

The *Su(var)3-9* protein family combines two of the most evolutionarily conserved domains of chromatin regulators: the chromo (Aasland, R. and Stewart, A.F., *Nucleic Acids Res 23*:3168-74 (1995); Koonin, E.V., *et al., Nucleic Acids Res 23*:4229-33 (1995)) and the SET (Jenuwein, T., *et al., Cell Mol Life Sci 54*:80-93 (1998)) domain. Whereas the 60 amino acid chromo domain represents an ancient histone-like fold (Ball, L.J., *et al., EMBO J 16*:2473-2481 (1997)) that directs eu- or heterochromatic localizations (Platero, J.S., *et al., Embo J 14*:3977-86 (1995)), the molecular role of the 130 amino acid SET domain has remained enigmatic.

The present invention started from the premise that the protein domain referred to as "SET" (Tschiersch, *et al., Embo J. 13(16)*:3822-3831 (1994))

defines a new genetic family of mammalian chromatin regulators which are important in terms of their developmental history on account of their evolutionary conservation and their presence in antagonistic gene products. Moreover, the characterization of other members of the group of SET domain genes, apart from *HRX/ALL-1*, helps to explain the mechanisms which are responsible for structural changes in chromatin possibly leading to malignant transformation.

One aspect of the present invention is therefore to identify mammalian, such as human and murine, chromatin regulator genes, clarify their function and use them for diagnosis and therapy. More specifically, the sequences of the *SUV39H* proteins, and variants thereof, and *EZH2* proteins, and variants thereof, according to the invention, may be used to analyze the interaction of SET domain proteins with chromatin or with other members of heterochromatin complexes. Starting from the findings thus obtained regarding the mode of activity of these proteins, the detailed possibilities for targeted intervention in the mechanisms involved therein are defined and may be used for therapeutic applications as described in detail below.

In order to achieve this objective, the sequence information of the SET domain was used to obtain the human cDNA homologous to the SET domain genes of *Drosophila* from human cDNA banks. Two cDNAs were obtained which constitute human homologues of *E(z)* and *Su(var)*3-9. The corresponding human genes are referred to as *EZH2* and *SUV39H*. *See* FIGS. 6 and 7. In addition, a variant form of *EZH2* was identified which is referred to as *EZH1*. *See* FIG. 8.

The present invention thus relates to DNA molecules containing a nucleotide sequence coding for a chromatin regulator protein which has a SET-domain, or a partial sequence thereof, characterized in that the nucleotide sequence is that shown in FIG. 6 (SEQ ID NO:1), or a partial sequence thereof, or FIG. 7 (SEQ ID NO:3), or a partial sequence thereof. The DNA molecules, including variants and mutants thereof such as dominant-negative mutants, are also referred to as "genes according to the invention." Two examples of genes

according to the invention are designated *EZH2* and *SUV39H*. They were originally referred to as "HEZ-2" and "H3-9," respectively.

According to another aspect, the invention relates to the cDNAs derived from the genes of the invention, including the degenerate variants thereof, and mutants thereof, which code for functional chromatin regulators and which can be traced back to gene duplication. An example of this is *EZH1* (SEQ ID NO:5), the partial sequence of which is shown by comparison with *EZH2* (SEQ ID NO:1) in FIG. 8.

According to another aspect, the invention relates to recombinant DNA molecules containing the cDNA molecules, functionally connected to expression control sequences, for expression in procaryotic or eucaryotic host organisms. Thus, the invention further relates to procaryotic or eucaryotic host organisms transformed with the recombinant DNA.

The invention further relates to antisense(deoxy)ribonucleotides with complementarity to a partial sequence of an inventive DNA molecule.

The invention further relates to transgenic animals, such as transgenic mice, which comprise a trans gene for the expression of a chromatin regulator gene which has a SET domain, or a mutated version or degenerate variant of such a protein.

The invention further relates to knock-out animals such as knock-out mice, obtainable from embryonic stem cells in which the endogenous mouse loci for *EZH1* and *SUV39H* are interrupted by homologous recombination.

The invention further relates to a process for identifying mammalian chromatin regulator genes which have a SET domain, or mutated versions thereof, wherein mammalian cDNA or genomic DNA libraries are hybridized under non-stringent conditions with a DNA molecule coding for the SET domain or a portion thereof.

The invention further relates to antibody molecules which bind to a polypeptide which contains the amino acid sequence depicted in SEQ ID NOS:2 or 4 or degenerate variants or mutants thereof.

Other aspects of the invention are set forth in the Detailed Description of the Preferred Embodiments.

Brief Description of the Figures

FIG. 1 is an amino acid sequence comparison between *EZH2* (SEQ ID NO:2) and *Drosophila enhancer of zeste (E(z))* (SEQ ID NO:11). The conserved carboxy terminal SET-domain (shaded box) and the Cys-rich region (Cys groups are emphasized) are shown. Percent identity is shown on the right side. The presumed nucleus locating signals are underlined.

FIG. 2 is an amino acid sequence comparison between the human homologue *SUV39H* (SEQ ID NO:4) and *Drosophila Su(var)3-9* (SEQ ID NO:16). The conserved carboxy terminal SET-domain (shaded box) and the Chromo-domain (darker shaded box) are shown. Percent identity is shown on the right side. The presumed nucleus locating signals are underlined. At the top of the figure is a diagrammatic summary of the two protein structures which shows that, in the human homologue, 207 amino acids are missing at the N-terminus.

FIG. 3 shows the aberrant transcripts of human SET-domain genes. On the left of the figure is the position of the five currently known SET-domain genes on the appropriate chromosome. The names of the authentic genes in each case are given on the right side of FIG. 3. More specifically, FIG. 3 shows, *inter alia*, the three genes (*HRX/ALL-1, EZH1/B52* and *SUV39H/MG-44*) for which aberrant cDNAs have been mapped on translocation fracture points or unstable chromatin regions. Four of the five SET-domain genes shown have mutations, all of which interrupt the carboxy terminal SET-domain. A translocation connects the amino terminal half of HRX to a non-correlated gene sequence which is shown as a dotted box designated ENL. Mutations and a premature stop codon change the SET-domain of *EZH1/B52*. Point and frameshift mutations interrupt the Chromo-

and SET-domain in *MG-44*. A large insertion cleaves the SET-domain of *KG-1* into two halves. At present, there are no known aberrant transcripts for *G9a*. The cysteine-rich cluster in *B52* is shown as a dotted box. In *HRX/ALL-1*, the region of homology with methyltransferase is shown as a shaded box and the A/T-hooks are shown as vertical lines.

FIG. 4 shows the evolutionary conservation of SET-domain proteins. Using the tfasta program of the Wisconsin GCG Network Service, proteins and open reading frames with homology to the SET-domain were identified. The figure shows a representative selection from yeasts to humans. The numbers indicate the amino acids. The carboxy terminal SET-domain is represented by a black box, Cys-rich regions are indicated by a darkly dotted box, and the chromo-domain of *Su(var)*3-9 and *SUV39H* are indicated by an open box with light dots. A region which is homologous to methyltransferase (*trx* and *HRX*) is shown as a shaded box. A/T hooks are indicated by vertical lines. Another Ser-rich region (S in C26E6.10) and a Glu-rich region (E in *G9a*) or ankyrin repeats (ANK in *G9a*) are also emphasized. YHR119 (GeneBank Accession No. U00059) and C26E6.10 (GeneBank Accession No. U13875) are open reading frames of cosmids in the databank without functional characterization. The percentages indicate the total amino acid identity between the human and the *Drosophila* proteins.

FIG. 5 shows the concordance between the amino acids in the SET domain in various *Drosophila* and human proteins. Specifically, the *EZH2* (SEQ ID NO:2) and *SUV39H* (SEQ ID NO:4) amino acid sequences were compared to the *E(z)* (SEQ ID NO:11), *HRX* (SEQ ID NO:12), *trx* (SEQ ID NO:13), *C26* (SEQ ID NO:14), *YHR* (SEQ ID NO:15), *Su(var)3-9* (SEQ ID NO:16); *G9a* (SEQ ID NO:17) and *KG-1* (SEQ ID NO:18) amino acid sequences. The SET domain of the genes shown in FIG. 5 was arranged using the Pileup program of the Wisconsin GCG Network Service. In order to compare the *KG-1* SET domain,

the large amino acid insert which splits the SET domain into two halves was removed before the pileup. *See* FIG. 3.

FIG. 6 illustrates the DNA and amino acid sequences of *EZH2* (SEQ ID NOS:1 and 2, respectively).

FIG. 7 illustrates the DNA and amino acid sequences of *SUV39H* (SEQ ID NOS:3 and 4, respectively).

FIG. 8 is a sequence comparison between the cDNAs of human *EZH2* (SEQ ID NO:1) and *EZH1* (SEQ ID NO:5). More specifically, FIG. 8 shows the nucleotide sequence of *EZH2* (SEQ ID NO:1) cDNA from position 1844 to 2330 in the upper line, the 5' splicing site and the potential stop codon being underlined. In order to ascribe a partial sequence of the cDNA of the *EZH1* variant (SEQ ID NO:5) to the *EZH2* sequence (SEQ ID NO:1) we used the gap program of the Wisconsin GCG Network Service. The premature stop codon in *EZH1* (position 353) is underlined. Sequences which code for the conserved SET-domain are emboldened. Moreover, the 3'-end (position 151 in *EZH1*) of the aberrant transcript *B52* (discussed below) is shown. Over the available sequence, *B52* was found to be 97% identical to *EZH1* and 72% identical to *EZH2*.

FIGS. 9A-B illustrate HMTase activity of transfected and recombinant *SUV39H1/Suv39h1* proteins. More specifically, in FIG. 9A, triple myc-tagged full-length human *SUV39H1* (aa 3-412) or a C-terminally truncated *SUV39H1* protein (aa 3-118) were immunoprecipitated from 'stably' transfected HeLa cell lines with anti-myc antibody beads and used in *in vitro* HMTase reactions with free histones as substrates and S-adenosyl-(*methyl*-^{14}C)-L-methionine as methyl donor. The Coomassie stain (top panel) shows purified proteins by arrowheads and free histones by dots. Fluorography (bottom panel) indicates HMTase activity of $(myc)_3$-*SUV39H1*(aa 3-412). In FIG. 9B, recombinant GST-fusion proteins

encoding different domains of murine *Suv39h1* were used in increasing protein concentrations for *in vitro* HMTase reactions as described above. The top panel is the Coomassie stain and the bottom panel is the fluorogram.

FIGS. 10A-C illustrate that lysine 9 of the H3 N-terminus is the major site for *in vitro* methylation by recombinant *Suv39h1*. More specifically, for FIG. 10A, approximately 10 μg of murine GST-*Suv39h1*(aa 82-412) were used in *in vitro* HMTase reactions with individual histones as outlined in FIGS. 9A-9B. The top panel is the Coomassie stain and the bottom panel is the fluorogram. For FIG. 10B, i*n vitro* methylation assays using GST-*Suv39h1*(aa 82-412) as enzyme and the indicated N-terminal peptides of wild-type H3, mutated H3 (K9L), CENP-A, macroH2A or insulin as substrates. FIG. 10C illustrates automated sequencing of the wild-type H3 N-terminal peptide (aa 1-20) that had been methylated *in vitro* by recombinant GST-*Suv39h1*(aa 82-412). Displayed is the ^{3}H-incorporation of individual amino acids identified at each successive round of microsequencing.

Detailed Description of the Preferred Embodiments

Sequencing

Starting from the sequence information of the conserved SET-domain, a human B-cell-specific cDNA library was screened, under reduced stringency, with a mixed *Drosophila*-DNA probe which codes for the SET-domains of *E(z)* and *Su(var)*3-9. From 500,000 plaques, 40 primary phages were selected. After another two rounds of screening, it became apparent that 31 phages code for authentic *E(z)*-sequences and 5 phages constitute *E(z)*-variants. By contrast, only two phages hybridized with the probe containing the SET-domain of *Su(var)*3-9 alone. The phage inserts were amplified by polymerase chain reaction (PCR) and analyzed by restriction mapping and partial sequencing. Representative cDNA inserts were subcloned and sequenced over their entire length. The 5'-ends were

isolated by screening positive phages once more with 5'-DNA probes, whereupon, after subcloning, complete cDNAs were obtained.

The complete cDNA coding for the human homologue of *E(z)* was designated *EZH2* (SEQ ID NO:1) and the DNA coding for the human homologue of *Su(var)*3-9 was designated *SUV39H* (SEQ ID NO:3). All in all, the identity of the amino acids between *Drosophila* and the human proteins amounts to 61% for *EZH2* and 43% for *SUV39H*, whilst the C-terminal SET-domain is very highly conserved (88% for *EZH2* and 53% *SUV39H*). Sequence comparison showed other clear regions of homology, e.g., a cysteine-rich domain in *EZH2* and a Chromo-Box in *SUV39H*. (In *polycomb*, it was shown that the Chromo-Box is the essential domain for the interaction between DNA and chromatin (Messmer, *et al., Genes & Dev. 6*:1241-1254 (1992))). By contrast, the 207 amino acids which make-up the amino terminal GTP-binding motif of the *Drosophila* protein are absent from the human homologue *SUV39H*. A comparison of the amino acid sequences between *Drosophila* and the human genes is shown in FIGS. 1 and 2. Moreover, another cDNA of the SET-domain family known as *MG-44* (see below) also lacks the 5'-end of the *Drosophila* gene.

Since translational consensus sequences are also present in the environment of the start-ATG of human *SUV39H*-cDNA, even at the corresponding internal position in *Su(var)*3-9, the *Drosophila* protein ought to contain additional exons which become dispensable for function at a later stage of evolution. The correctness of this hypothesis can be confirmed by expressing human *SUV39H*-cDNA and cDNAs of *Su(var)*3-9 which are either complete or shortened at the 5'-end in *Drosophila*.

In addition to the human cDNA of *SUV39H*, the homologous locus was also isolated in the mouse, the sequence analysis and promoter structure of which clearly confirm the amino terminal shortening of mammal-homologous genes compared with *Drosophila Su(var)3-9*.

DNA blot analyses carried out within the scope of the present invention indicate that mammal-homologous genes of *Su(var)3-9* are represented in mice

and humans by individual loci, whereas mammal-homologous genes of *E(z)* are coded by two separate loci in mice and humans. The second human locus (known as *EZH1*) was confirmed by characterizing a small number of cDNA variants which differ in their 3'-flanking sequences from the majority of the clones isolated from the human cDNA library. The differences between *EZH2* (SEQ ID NO:1) and *EZH1* (SEQ ID NO:5) in the sequenced area are shown in FIG. 8. The SET-domain of *EZH1* exhibits mutations compared with *EZH2*. Moreover, the *EZH1* variant which was isolated (in all probability, an aberrantly spliced cDNA) carries a stop codon located in the reading frame which shortens the protein by 47 C-terminal amino acids. Sequence comparison of *EZH1* (SEQ ID NO:5) with *EZH2* (SEQ ID NO:1) and the finding that there are two separate *E(z)*-homologous loci in humans and in mice, lead one to conclude that gene duplication has occurred in mammals.

In the light of the knowledge of the nucleotide sequence of the SET domain genes, it is possible to produce the corresponding proteins derived from the cDNA sequences, which is also an object of the present invention, in recombinant form, by inserting the cDNAs coding for them in suitable vectors and expressing them in host organisms. The techniques used to produce recombinant proteins are well known to the skilled person and may be taken from relevant manuals (Sambrook, J., Fritsch, E.F. and Maniatis, T., 1989, Cold Spring Harbor Laboratory Press). The present invention thus relates, in another aspect, to recombinant DNA molecules, containing the DNA coding for *EZH2* (SEQ ID NO:1) or variants thereof, *SUV39H* (SEQ ID NO:3) or variants thereof, or *EZH1* (SEQ ID NO:5) or variants thereof, or another SET-dependent protein or variant thereof, expression control sequences functionally connected thereto, and the host organisms transformed therewith.

SET Domain Mutations and Functionality

In a comparison with cDNA sequences in the GeneBank databank, it was surprisingly found that certain cDNA partial sequences recorded in the databank, which are derived from aberrant transcripts in tumor tissues, constitute mutated versions of the cDNAs according to the invention. For example, in the search for *BRCA1*, a gene which indicates a predisposition to breast and ovarian cancer, a partial cDNA sequence with 271 nucleotides was isolated, known as *B52*, which codes for a mutated variant of the SET-domain and it was mapped on the human chromosome 17q21 (Friedman, *et al., Cancer Research 54*:6374-6382 (1994)). Within the scope of the present invention, it was surprisingly found that *B52* shows 97% identity with the *EZH1* cDNA variant according to the invention. *EZH1* might possibly be a gene the reactivation of which plays a part in deregulated proliferation.

As another example, a cDNA (2,800 nucleotides; *MG-44*) was isolated from human chromosome Xp11 (Geraghty, *et al., Genomics 16*:440-446 (1993)), a region which indicates a predisposition to degenerative disorders of the retina and synovial sarcoma. It was found, surprisingly, that this cDNA has 98% identity with the *SUV39H* cDNA according to the invention.

The new genes prepared within the scope of the present invention thus make it possible to infer a correlation between certain cancers and mutations in chromatin regulators. For example, in the case of *MG-44* cDNA, as it has numerous point and frameshift mutations which interrupt the chromo- and SET-domains, it became possible for the first time, using the *SUV39H* cDNA according to the invention, to clarify a correlation between *Su(var)*3-9 and *MG-44*.

Apart from the sequences already mentioned, the GeneBank databank also records, as other human members of the SET-protein family, the well-documented human homologue of *Drosophila trx, HRX/ALL-1* (Tkachuk, *et al., Cell 71*:691-700 (1992); Gu, *et al., Cell 71*:701-708 (1992)); a gene of unknown function known as *G9a* which is present in the human Major Histocompatibility Complex

(Milner and Campbell, *Biochem J. 290*:811-818 (1993)); and thirdly, an unpublished cDNA (*KG-1*) which was isolated from immature myeloid tumor cells (Nomura, *et al.*, Unpublished, GeneBank Accession Number:D31891 (1994)). Whereas *G9a* is currently the only human gene with a SET-domain for which no mutated version is known hitherto, *KG-1* carries an insertion of 342 amino acids which cleaves the SET-domain into an amino-terminal half and a carboxy-terminal half. Probably, this *KG-1* cDNA constitutes an aberrantly spliced variant since there are 5' and 3' consensus splicing sites at both ends of the insertion. In all, four of the five currently known human members of the SET-protein family have undergone changes, all of which mutate the SET-domain (*HRX/ALL-1, EZH1/B52, SUV39H/MG-44* and *KG-1*). Moreover, in three cases, the corresponding human gene loci in the vicinity of translocational fracture points or unstable chromosomal regions have been mapped (*HRX/ALL-1, EZH1/B52* and *SUV39H/MG-44*). *See* FIG. 3.

The fact that a mammalian gene of the SET-protein family, *HRX/ALL-1*, has been connected with translocation-induced leukaemogenesis (Tkachuk, *et al.*, *Cell 71*:691-700 (1992); Gu, *et al.*, *Cell 71*:701-708 (1992)) is a strong indication that proteins with the SET-domain are not only important regulators of development which co-determine chromatin-dependent changes in gene expression, but that, after mutation, they also disrupt normal cell proliferation.

Since all the mutations described hitherto interrupt the primary structure of the SET-domain, it is fair to assume that it is the SET-domain as such which plays a crucial part in the transition from the normal state into the transformed state. Furthermore, the SET-domain may have an important role in view of its evolutionary conservation in gene products which occurs from yeasts to humans.

To investigate the frequency with which the SET domain is subjected to specific mutations, it is possible to use the SET-specific DNA probes to analyze single-strand conformation polymorphisms (SSCP; Gibbons, *et al.*, *Cell 80*:837-845 (1995)). Types of cancer in which SET-specific DNA probes can be used as diagnostic markers are breast cancer (*EZH1*; Friedman, *et al.*, *Cancer Research*

54:6374-6382 (1994)), synovial sarcoma (*SUV39H*; Geraghty, *et al., Genomics 16*:440-446 (1993)) and leukaemias.

It has been assumed by other authors (DeCamillis, *et al., Genes & Dev. 6*:223-232 (1992); Rastelli, *et al., Embo J. 12*:1513-1522 (1993); Orlando and Paro, *Cell 75*:1187-1198 (1993)) that complexing between various members of heterochromatin proteins is essential for their functioning. In view of the availability of the SET domain genes according to the invention, it is possible to determine whether the SET region constitutes a domain which functions because of interactions or whether it contributes to the formation of multimeric heterochromatic complexes. Similarly, it is possible to determine whether the SET domain has an inhibitory function, similar to the amino-terminal BTB domain of various chromatin regulators, including the GAGA factor (Adams, *et al., Genes & Dev. 6*:1589-1607 (1992)).

Investigations which serve to analyze the function of the SET domain may be carried out, for example, by expressing cDNAs coding for human *EZH2* or *SUV39H*, and providing an epitope against which antibodies are available *in vitro* and in tissue cultures. After immune precipitation with the appropriate epitope-specific antibodies, it is possible to establish whether *EZH2* and *SUV39H* are able to interact with each other *in vitro* and whether complexing occurs *in vivo* between *EZH2* and/or *SUV39H* with other chromatin regulators. In all, the analyses of interactions with *EZH2* and *SUV39H* proteins provided with epitopes allow for further characterization of the function of the SET domain. This opens up possibilities of taking action against deregulated activity by, e.g., introducing dominant-negative variants of the SET domain cDNA sequences into the cell using gene-therapy methods. Such variants are obtained, for example, by first defining the functional domains of the SET proteins, e.g., the sequence portions responsible for the DNA/chromatin interaction or protein/protein interaction, and then expressing the DNA sequences shortened by the relevant domain(s), or sections thereof, in the cell in question in order to compete with the deregulated proliferation caused by the intact functional protein.

The availability of the cDNAs according to the invention also makes it possible to produce transgenic animals, e.g., mice, wherein SET domain genes can either be overexpressed ("gain-of-function") or wherein these genes can be switched off ("loss-of-function"). Such transgenic animals are also an object of the present invention.

In particular, the "gain-of-function" analyses, in which alleles of the genes according to the invention are introduced into the mouse, provide final conclusions as to the causative participation of *EZH2* and *SUV39H* in the chromatin-dependent requirements of tumor formation. For the "gain-of-function" analysis, the complete cDNA sequences of human *EZH2* and *SUV39H,* and mutated versions thereof, such as *EZH1/B52* and *MG-44*, may be driven by vectors which allow high expression rates, e.g., plasmids with the human β-actin promoter, and by the enhancer of the heavy chain of immunoglobulins (Eμ) and also by Moloney virus enhancers (Mo-LTR). Recently, it was shown that the Eμ/Mo-LTR-dependent overexpression of the *bmi* gene, which, in common with *EZH2*, belongs to the *Pc* group of negative chromatin regulators, is sufficient to produce lymphomas in transgenic mice (Alkema, *et al., Nature 374*:724-727 (1995)).

In the "loss-of-function" analyses, the endogenous mouse loci for *EZH1* and *SUV39H* are interrupted by homologous recombination in embryonic stem cells, thus, it is possible to determine whether the loss of the *in vivo* gene function leads to abnormal development of the mouse.

As a result of these *in vivo* systems, the activity of *EZH2* and *SUV39H* can be confirmed. These systems also form the basis for animal models in connection with human gene therapy.

For a detailed analysis of the function of the cDNAs according to the invention or partial sequences thereof with respect to the diagnostic use of SET domain gene sequences, within the scope of the present invention, homologous murine cDNAs were isolated from *EZH1* and *SUV39H*. When using a mouse-specific DNA probe coding for the SET domain in "RNAse protection" analyses

to investigate the *EZH1* gene activity during normal mouse development, a somewhat broad expression profile became apparent which is similar to that of the *bmi* gene (Haupt, *et al., Cell 65*:753-763 (1991)). The analyses carried out with the murine sequences were expanded with human sequences to compare the quantities of RNA between immature precursor cells, tumor cells and differentiated cells in various human cell culture systems.

SUV39H Proteins

Overexpression studies with human *SUV39H* mutants indicate a dominant interference with higher-order chromatin organization that, surprisingly, suggests a functional relationship between the SET domain and the distribution of phosphorylated (at serine 10) H3 (Melcher, M., *et al., Mol Cell Biol 20*:3728-41 (2000)). The experiments of the present invention, as shown in the Examples, show that mammalian *SUV39H1*, or other *SUV39H* proteins, are SET domain-dependent, H3-specific histone methyltransferases (HMTases) which selectively methylate lysine 9 of the H3 N-terminus. *See* FIGS. 9 and 10. Methylation of lysine 9 negatively regulates phosphorylation of serine 10 and reveals a histone code that appears intrinsically linked to the organization of higher-order chromatin.

In the present invention, the function of members of the SU(VAR)3-9 protein family was investigated with the view to develop novel strategies to interfere with chromosome stability and high fidelity chromosome segregation. Such strategies can be employed in therapies for the treatment of conditions in which aberrant gene expression and genomic instability through chromosome missegregation are causally involved. (The term "high fidelity chromosome segregation" implies successful segregation of chromosomes resulting in the maintenance of a stable karyotype).

To this end, in a first step, bioinformatic techniques were applied. Using the SET domains of the SU(VAR)3-9 protein family as a starting alignment,

distant sequence and secondary structure similarities to six plant protein methyltransferases were detected. To investigate whether the SET domain of human *SUV39H1* has enzymatic activity, histones were tested as possible substrates for *in vitro* methylation. The obtained results demonstrate that *SUV39H1* harbors an intrinsic histone methyltransferase activity and suggest that this HMTase activity resides in the C-terminal SET domain. Experiments indicated that the HMTase activity of mammalian SU(VAR)3-9 related proteins is selective for H3 under the chosen assay conditions. To examine this finding in more detail, *in vitro* methylation reactions were performed with individual histones. It could be shown that H3 is specifically methylated by GST-*Suv39h1*(aa 82-412), whereas no signals are detected with H2A, H2B or H4. Methylation of H3 has been shown to occur predominantly at lysine 4 in a wide range of organisms, as well as at lysine 9 in HeLa cells, although the responsible HMTase(s) have yet to be defined. To investigate the site utilization profile of *Suv39h1*, unmodified peptides comprising the wild-type H3 N-terminus and a mutant K9L peptide were tested as substrates. Additionally, insulin and peptides comprising the N-termini of CENP-A and macroH2A were included. These *in vitro* assays revealed selective methylation of the wild-type H3 peptide. The data obtained also suggested that the H3 N-terminus is a preferred residue for *Suv39h1*-dependent HMTase activity. To more definitively determine this site preference, the wild-type H3 N-terminal peptide was *in vitro* methylated by GST-*Suv39h1*(aa 82-412), using S-adenosyl-(*methyl*-^{3}H)-L-methionine. The labeled peptide, purified by reverse-phase HPLC, was then directly microsequenced, and ^{3}H-incorporation associated with each individual amino acid was analyzed. The results confirmed selective transfer of methyl-label to lysine 9, demonstrating that *Suv39h1* is a highly site-specific HMTase for the H3 N-terminus *in vitro* (FIG. 10C). The identification of members of the SU(VAR)3-9 protein family, exemplified by human *SUV39H1*, murine *Suv39h1* and murine *Suv39h2*, as K9 specific histone H3 MTases is the prerequisite for designing assay methods that allow for finding compounds altering, in particular

interfering with, chromosome stability, which is the basis for novel therapeutic approaches. *Suv39h* proteins and other methyl transferases with *Suv39h*-like activity are useful in a method for identifying compounds that have the ability of modulating chromosome stability in plant or animal cells. This method is characterized in that a MTase with *Suv39h*-like MTase activity is incubated, in the presence of the substrate(s) for its enzyme activity and optionally its co-factor(s), with test compounds and that the modulating effect of the test compounds on the MTase activity of the MTase is determined.

Since it has been shown in the present invention that recombinant *Suv39h* retains MTase activity, most preferably, recombinant enzymes are employed. *Suv39h* or *Suv39h* variants can be produced recombinantly according to standard methods by expression in suitable hosts, e.g., bacteria, yeast, insect or eucaryotic cells and purified, e.g., on glutathione-agarose columns if it has been tagged with GST. For testing the compounds for their effect on *Suv39h* activity, the assay comprises, as its essential features, incubating a histone H3 protein or histone H3 N-terminal fragment including K9, a methyl donor, S-adenosyl-L-Methionine with a preparation containing a *Suv39h* MTase activity and determining MTase activity in the presence or absence of a test substance.

MTase substrates useful in the method of the invention may be those equivalent to or mimicking the naturally occurring substrates, e.g., biochemically purified histone H3, recombinantly produced histone H3, or a histone H3 peptide that contains the K9 methylation site, or other yet to be identified proteins which act as substrates for *Suv39h* MTases. Additional novel *Suv39h* substrates can be identified by bioinformatic/biochemical techniques and tested using the biochemical assays described herein. These novel *Suv39h* substrates can be identified by co-immunoprecipitation techniques. *Suv39h* proteins or tagged versions of *Suv39h* proteins could be immunoprecipitated with specific anti-sera and interacting proteins identified by mass spectroscopy techniques. A yeast two-hybrid screen using *Suv39h* proteins or portions of *Suv39h* proteins as a bait could

also be employed to identify novel interacting protein from a variety of cDNA libraries.

In a preferred embodiment, the histone H3 fragment ARTKQTARKSTGGKAPRKQL (SEQ ID NO:19) is employed. Alternatively, a similar peptide may be used for which the MTase has increased affinity/activity. The methyl donor preferably carries a detectable label, e.g., a radioactive or a chromogenic label, which can be quantified upon transfer to the substrate. Preferably, the methyl donor is the natural methyl donor S-adenosyl-L-Methionine. Alternatively to using a labeled methyl donor, the substrate, upon methylation by the enzyme, serves as an epitope which can be recognized by a specific antibody and hence used for quantification by standard immunoassay techniques, e.g., ELISAs. Antibodies useful in this type of assay can be obtained by using the methylated substrate, preferably a small peptide, e.g., the K9 methylated peptide ARTKQTARKSTGGKAPRKQL (SEQ ID NO:19) as an antigen and obtaining polyclonal or monoclonal antibodies according to standard techniques. For small scale applications, the screening method can be based on the principal of the assay as described in Example 3. In a preferred embodiment, the method is performed on a high-throughput scale. For this embodiment the major assay components, in particular *Suv39h*, are employed in recombinant form. The thus obtained recombinant protein can then be used in an inhibitor screen. For the high-throughput format, the screening methods to identify MTase inhibitors, are carried out according to standard assay procedures. Such assays are based on the catalytic transfer, mediated by *Suv39h* or a *Suv39h* variant, of a methyl group from a substrate to a histone H3 peptide. To achieve this, the substrate histone H3 peptide would be immobilized and incubated with recombinant *Suv39h* or *Suv39h* variant and a chromogenic methyl donor or radioactively labeled methyl donor or a unmodified methyl donor. Upon transfer of the methyl group to the histone H3 peptide by *Suv39h*, the chromogenic methyl donor would change color which and can be quantified or the radioactive methyl group transferred to the substrate quantified or the methylation of the substrate

quantified by ELISA using an antibody specific for the methylated substrate. If a test substance is an inhibitor of the MTase activity, there will be, depending on the detection system and depending on whether the test substance has an inhibiting or an activating effect, a decrease or an increase in the detectable signal. In the high-throughput format, compounds with a modulating effect *Suv39h* MTase activity can be identified by screening test substances from compound libraries according to known assay principles, e.g., in an automated system on microtiter plates.

Applications for Therapy

On the basis of the criteria laid down within the scope of the present invention, it transpires that the genes which have a SET domain are involved in the chromatin-dependent occurrence of deregulated proliferation. These genes or the cDNAs derived therefrom, or partial or mutated sequences thereof, can thus be used in the treatment and diagnosis of diseases which can be attributed to such proliferation. Specifically, oligonucleotides coding for the SET domain as such or parts thereof may be used as diagnostic markers in order to diagnose certain types of cancer in which the SET domain is mutated.

The DNA sequences according to the invention, or sequences derived therefrom, e.g., complementary antisense oligonucleotides, may be used in gene therapy - depending on whether the disease to be treated can be put down to deregulation of chromatin as a result of the absence of the functional gene sequence or as a result of overexpression of the corresponding gene(s) - by introducing the functional gene sequence, by inhibiting gene expression, e.g., using antisense oligonucleotides, or by introducing a sequence coding for a dominant-negative mutant. For example, as *SUV39H* is required to maintain a stable karyotype as described above, it can be considered as possessing tumor suppressor gene activity. If *SUV39H* mutations are factors underlying cellular transformation

events, the re-introduction of a wild type *SUV39H* gene by gene therapy may result in increased genomic stability delaying or inhibiting cancer progression.

The inventive DNA molecules may be administered, preferably in recombinant form as plasmids, directly or as part of a recombinant virus or bacterium. In theory, any method of gene therapy may be used for therapy of cancer based on DNA, e.g., on *SUV39H* DNA, both *in vivo* and *ex vivo*. Thus, the DNA sequences in question may be inserted into the cell using standard processes for the transfection of higher eukaryotic cells, which may include gene transfer using viral vectors (retrovirus, adenovirus, adeno-associated virus, vaccinia virus or *Listeria monocytogenes*) or using non-viral systems based on receptor-mediated endocytosis. Surveys of the common methods are provided by, for example, Mitani, K. and Caskey, C.T., *Trends in Biotechnology 11*:162-166; Jolly, D., *Cancer Gene Therapy 1*:51 (1994); Vile, R. and Russel, S., *Gene Therapy 1*:88 (1994); Tepper, R.I. and Mule, J.J., *Human Gene Therapy 5*:153 (1994); Zatloukal, K, *et al., Gene 135*:199 (1993); WO 93/07283. Examples of *in vivo* administration are the direct injection of "naked" DNA, either by intramuscular route or using a gene gun. Moreover, synthetic carriers for nucleic acids such as cationic lipids, microspheres, micropellets or liposomes may be used for *in vivo* administration of nucleic acid molecules coding for the *SUV39H* polypeptide.

To inhibit the expression of the genes according to the invention, it is also possible to use lower-molecular substances which interfere with the machinery of transcription. After analyzing the 5'-regulatory region of the genes, it is possible to screen for substances which wholly or partially block the interaction of the relevant transcription factors with this region by, e.g., using the method described in WO 92/13092.

Inhibition of deregulated proliferation may also act on the gene product, by therapeutically using the corresponding antibodies against the *EZH2*- or *SUV39H*- protein, preferably human or humanized antibodies. Such antibodies are produced by known methods, e.g., as described by Malavsi, F. and Albertini, A.,

TIBTECH 10:267-269 (1992), or by Rhein, R., *The Journal of NIH Res. 5*:40-46 (1993). Thus, the invention also relates to antibodies against *EZH2* or *SUV39H* or other SET-dependent proteins which may be used therapeutically or diagnostically.

As another therapeutic approach, by providing a method to identify compounds which exert their effect by directly modulating, in particular, by inhibiting, *SUV39H*, for example, a novel approach for inhibiting the proliferation of rapidly dividing animal cells, in particular tumor cells, is provided. Compounds identified in the above-described assays have the ability to modulate chromosome stability by modulating the MTase activity of *SUV39H*. Compounds, which act as modulators of *SUV39H*, can also be used in human therapy, in particular cancer therapy.

The efficacy of compounds identified as *SUV39H* modulators can be tested for *in vivo* efficacy in mammalian cells with *SUV39H* double null cells serving as a positive control. Effective compounds should interfere with chromosome stability which can be measured by karyotyping, e.g., by analyzing DNA content by FACS, or by standard cytological techniques. Substances whose potential for therapeutic use has been confirmed in such secondary screen can be further tested for their effect on tumor cells.

To test the inhibition of tumor cell proliferation, primary human tumor cells may be incubated with the compound identified in the screen and the inhibition of tumor cell proliferation tested by conventional methods, e.g., bromo-desoxy-uridine or ^{3}H incorporation. Compounds that exhibit an anti-proliferative effect in these assays may be further tested in tumor animal models and used for the therapy of tumors.

Toxicity and therapeutic efficacy of the compounds identified as drug candidates by the methods described above can be determined by standard pharmaceutical procedures, which include conducting cell culture and animal experiments to determine the IC_{50}, LD_50 and ED_{50}. The data obtained may be used for determining the human dose range, which will also depend on the dosage

form (tablets, capsules, aerosol sprays, ampules, etc.) and the administration route (oral, buccal, nasal, parenteral, rectal, etc.). A pharmaceutical composition containing the compound as the active ingredient may be formulated in a conventional manner using one or more physiologically active carriers and excipients. Methods for making such formulations can be found in manuals, e.g., "Remington Pharmaceutical Sciences."

SUV39H mediates dynamic transitions in higher order mammalian chromatin in part through its intrinsic HMTase activity. K9 methylation of histone H3 (K9-Me) represents an important epigenetic imprint for chromosome dynamics during cell division. Antibodies specific for K9-Me could be used to screen cells/patients for heterochromatin based genome instabilities. In essence, K9-Me specific anti-sera can be used a diagnostic tool for several potential human diseases.

Further, differences in the transcription level of SET domain RNAs between normal and transformed cells can be used as diagnostic parameters for diseases in which the expression of SET domain genes is deregulated. To find out whether the SET domain is accordingly suitable as a diagnostic tumor marker for specific cancers or as a general diagnostic parameter, it is possible to use current methods for determining the RNA concentration, as described in the relevant laboratory manuals (Sambrook, J., Fritsch, E.F. and Maniatis, T., 1989, Cold Spring Harbor Laboratory Press) such as Northern Blot, S1-nuclease protection analysis or RNAse protection analysis.

The following examples are provided by way of illustration to further describe certain preferred embodiments of the invention, and are not intended to be limiting of the present invention, unless specified.

Examples

Example 1

Preparation of a cDNA library

Human B-cell-specific cDNA library as described by Bardwell and Treisman, *Genes & Dev. 8*:1644-1677 (1994), was prepared by isolating poly(A)$^+$-RNA from human BJA-B-cells, reverse-transcribing it by poly(dT)$_{15}$, priming and converting it into double-stranded cDNA. After the addition of an EcoRI adapter of the sequence 5' AATTCTCGAGCTCGTCGACA (SEQ ID NO:6), the cDNA was ligated into the EcoRI site of the bacteriophage gt10. The propagation and amplificiation of the library were carried out in *E. coli* C600.

Preparation of DNA probes

Drosophila DNA probes coding for the conserved SET domains of *E(z)* and *Su(var)*3-9 were prepared on the basis of the published *Drosophila* sequences (Jones and Gelbart, *MCB 13(10)*:6357-6366 (1993); Tschiersch, *et al., Embo J. 13(16)*:3822-3831 (1994)) by polymerase chain reaction (PCR): 1 µg of *Drosophila melanogaster*-DNA (Clontech) was subjected with the two primers, *E(z)* 1910 (5'ACTGAATTCGGCTGGGGCATCTTTCTTAAGG) (SEQ ID NO:7) and *E(z)* 2280 (5' ACTCTAGACAATTTCCATTTCACGCTCTATG) (SEQ ID NO:8), to PCR amplification (35 cycles of 30 sec at 94°C, 30 sec at 55°C and 30 sec at 72°C). The corresponding SET domain probe for *Su(var)*3-9 was amplified from 10 ng of plasmid DNA (Tschiersch et al., 1994; clone M4)

with the pair of primers suvar.up (5' ATATAGTACTTCAAGTCCATTCAAAAGAGG) (SEQ ID NO:9) and suvar.dn (5' CCAGGTACCGTTGGTGCTGTTTAAGACCG) (SEQ ID NO:10), using the same cycle conditions. The SET domain DNA fragments obtained were gel-purified and partially sequenced in order to verify the accuracy of the amplified sequences.

Screening the cDNA library

5×10^5 plaque forming units (pfu) were incubated with 5 ml of culture of the bacterial host strain of *E. coli* C600 (suspended at an optical density OD_{600} of 0.5 in 10 mM $MgSO_4$) at 37°C for 15 min and then poured onto a large (200 mm x 200 mm) preheated LB dish. After growing overnight at 37°C, the phages were absorbed on a nylon membrane (GeneScreen). The membrane was left floating with the side containing the absorbed phages facing upwards, for 30 sec in denaturing solution (1.5 M NaCl, 0.5 M NaOH), then immersed for 60 sec in denaturing solution and finally neutralized for 5 min in 3 M NaCl, 0.5 M Tris (pH 8). The membrane was then briefly rinsed in 3xSSC and the phage DNA was fixed on the nylon filter by UV-crosslinking. The filter was prehybridized for 30 min at 50°C in 30 ml of Church buffer (1 % BSA, 1 mM EDTA and 0.5 M $NaHPO_4$, pH 7.2), then 2×10^6 cpm of the radiolabeled DNA probe mixture of *E(z)*-SET and *Su(var)3-9*-SET were added. The DNA probes were prepared by random priming using the RediPrime Kit (Amersham). Hybridization was carried out overnight at 50°C. After the hybridizing solution had been removed, the filter was washed for 10 sec in 2xSSC, 1 % SDS at ambient temperature, then for 10 sec at 50°C. The filter was wrapped in Saranwrap and subjected to autoradiography using an intensifier film.

Positive phage colonies were identified on the original plate by matching the autoradiogram and the corresponding agar fragments were removed using the larger end of a Pasteur pipette. The phage pool was eluted overnight at 4°C in 1

ml SM-Buffer (5.8 g NaCl, 2 g $MgSO_4$-H_2O, 50 ml Tris (pH 7.5), 5 ml 2% gelatine on 1 l H_2O), containing a few drops of $CHCl_3$. The phage lysate was plated out for a second and third round of screening in order to obtain individual, well isolated positive plaques (20 to 100 plaques per plate in the third round).

Sequence analysis

The cDNA inserts from recombinant phages were subcloned into the polylinker of pBluescript KS (Stratagene) and sequenced in an automatic sequencer (Applied Biosystems) using the dideoxy method. The complete sequence of at least two independent isolates per gene obtained was determined by primer walking. The sequences were analyzed with the GCG-Software package (University of Wisconsin), and the investigation for homology was carried out using the "Blast and fasta" or "tfasta" network service. The complete sequences of *EZH2* (SEQ ID NO:1) and *SUV39H* (SEQ ID NO:3) are shown in FIGS. 6 and 7.

Examples 2-4

Materials and Methods

Sequence alignments and secondary structure predictions

The SET domains of human *SUV39H1*, *Drosophila Su(var)3-9* and *S. pombe* CLR4 were used as a multiple starting alignment for database similarity searches using Profile, hidden Markov and position-specific iterative BLAST methods (representative listings are available from the SET domain page of the SMART WWW-server). These searches revealed significant similarities to six plant proteins (accession numbers Q43088, O65218, P94026, O80013, AAC29137 and AC007576_12) described as putative lysine N-methyltransferases.

For example, a PSI-BLAST search with the *S. pombe* hypothetical protein SPAC3c7.09 as query identified these plant sequences and well-known SET domain sequences within ten rounds using an E-value inclusion threshold of 0.001. The same search also revealed the presence of a SET domain in YHR109w (which is known to encode a cytochrome c MTase) within three rounds. Consensus secondary structures were predicted by described algorithms.

Epitope-tagged SUV39H1 proteins in HeLa cells

The HeLa cell lines overexpressing full-length $(myc)_3$-*SUV39H1* (aa 3-412) or $(myc)_3$-Nchromo (aa 3-118) have been described. Nuclear extracts were immunoprecipitated with anti-myc antibody beads, and approximately 1-3 μg of matrix-bound $(myc)_3$-tagged *SUV39H1* proteins were used for *in vitro* HMTase assays.

Generation and purification of GST-fusion proteins

The GST-*Suv1*(aa 82-412) product expressed from the pGEX-2T vector (Pharmacia) as a glutathione-*S*-transferase (GST) fusion protein has been described. Additional GST constructs were generated by transferring BamHI-EcoRI PCR amplicons into pGEX-2T. All constructs were confirmed by sequencing. Recombinant proteins were expressed in 1l cultures of *E. coli* strain BL21 and solubilized in 10 ml RIPA buffer ((20 mM Tris (pH 7.5), 500 mM NaCl, 5 mM EDTA, 1% NP-40, 0.5% sodium deoxycholate) containing a full set of protease inhibitors (Boehringer Mannheim) and lysozyme (5 mg/ml; Sigma)) by freeze-thawing in liquid N_2, followed by sonication. Soluble proteins were cleared by centrifugation, purified with 800 ml glutathione Sepharose beads (Pharmacia) and washed twice in RIPA buffer. Protein concentration was determined by Coomassie staining of SDS-PAGE gels. Matrix-bound fusion proteins were used immediately for *in vitro* HMTase assays or stored at 4°C.

In vitro histone methyltransferase (HMTase) assay

In vitro HMTase reactions were modified based on described protocols and carried out in a volume of 50 μl of methylase activity buffer (MAB: 50 mM Tris (pH 8.5), 20 mM KCl, 10 mM $MgCl_2$, 10 mM b-ME, 250 mM sucrose), containing 10 μg of free histones (mixture of H1, H3, H2B, H2A and H4; Boehringer Mannheim) as substrates and 300 nCi S-adenosyl-(*methyl*-^{14}C)-L-methionine (25 mCi/ml) (Amersham) as methyl donor. 10 μg of matrix-bound GST-fusion proteins were routinely used to assay for HMTase activity. After incubation for 60 min. at 37°C, reactions were stopped by boiling in SDS loading buffer, and proteins were separated by 15% or 18% SDS-PAGE and visualized by Coomassie staining and fluorography. HMTase assays with individual histones (Boehringer Mannheim), insulin (Sigma) or N-terminal peptides were performed with 5 μg of substrate. The following peptides were used: wild-type N-terminus of human histone H3 (ARTKQTARKSTGGKAPRKQL) (SEQ ID NO:19) and mutant peptide which changes lysine 9 (bold) to leucine; N-terminus of human CENP-A (MGPRRRSRKPEAPRRRSPSP) (SEQ ID NO:20); N-terminus of rat macro-H2A (MSSRGGKKKSTKTSRSAKAG) (SEQ ID NO:21). Peptide microsequencing of the *in vitro* methylated wild-type H3 N-terminal peptide and determination of ^{3}H-incorporation of individual amino acids by scintillation counting was done as described.

Example 2

Sequence similarity of SET domains with plant methyltransferases

Using the SET domains of the SU(VAR)3-9 protein family as a starting alignment, significant sequence and secondary structure similarities (*see* Methods above) to six plant protein methyltransferases were detected. Although some of these plant sequences have been classified as potential histone lysine N-methyltransferases, only one had been functionally characterized, but was found to lack HMTase activity. Detected were amino acid and secondary structure (β-sheet (b) or α-helix (h)) similarities of the C-terminal halves of SET domain sequences from human *SUV39H1* (AF019968), murine *Suv39h1* (AF019969), murine *Suv39h2* (AF149205), *Drosophila Su(var)3-9* (P45975), a *C. elegans Su(var)3-9*-like ORF C15H11.5 (CAB02737), *S. pombe* CLR4 (O74565), human *EZH2* (Q15910), the human trithorax homologue HRX (Q03164), and MTases from *P. sativum* (rubisco ls-MT; Q43088) and *A. thaliana* (O65218). The plant MTase sequences contain an insertion of approximately 100 amino acids in the middle of the SET domain.

Example 3

HMTase activity of transfected and recombinant SUV39H1 and Suv39h1 proteins

To investigate whether the SET domain of human *SUV39H1* has enzymatic activity, histones were tested as possible substrates for *in vitro* methylation. Using HeLa cell lines 'stably' expressing triple myc-tagged full-length *SUV39H1* (aa 3-412), the ectopic protein was enriched from nuclear extracts by immunoprecipitation with anti-myc beads (*see* FIG. 9A, arrowhead top panel) and probed for activity to transfer a labeled methyl group from S-adenosyl-(*methyl*-^{14}C)-L-methionine to free histones according to described

conditions. Reaction products were separated by SDS-PAGE and visualized by fluorography, indicating selective transfer of the methyl-label to H3 (FIG. 9A, bottom panel). By contrast, no signals were detected with extracts from a HeLa cell line that expresses only the N-terminal third of *SUV39H1* (aa 3-118) or with extracts from HeLa control cells. To confirm that the HMTase activity is an intrinsic property of *SUV39H1* and not mediated by a *SUV39H1*-associated factor, the *in vitro* HMTase reactions was repeated with recombinant products that were purified as GST-fusion proteins from *E. coli* (*see* FIG. 9B, arrowheads top panel). For this analysis, murine *Suv39h1*, which is 95% identical to human *SUV39H1* (Aagaard, L., *et al., EMBO J. 18*:1923-1938 (1999)) was used. A purified GST-product comprising aa 82-412 maintained HMTase activity (although at a reduced level as compared to transfected *SUV39H1*), whereas a purified GST-product comprising aa 7-221 proved negative, even at higher protein concentrations (FIG. 9B, bottom panel). These results suggest that the HMTase activity resides in the C-terminal SET domain.

Example 4

Lysine 9 of the H3 N-terminus is the major site for in vitro methylation by recombinant Suv39h1.

The above Examples indicated that the HMTase activity of mammalian *Su(var)3-9* related proteins is selective for H3 under the chosen assay conditions. To examine this finding in more detail, *in vitro* methylation reactions were performed with individual histones, using GST-*Suv39h1*(aa 82-412) as an enzyme. As shown in FIG. 10A, H3 is specifically methylated by GST-*Suv39h1*(aa 82-412), whereas no signals are detected with H2A, H2B or H4. A weak signal is present if H1 was used as the sole substrate; the significance of H1 methylation remains to be determined. Methylation of H3 has been shown to occur predominantly at lysine 4 in a wide range of organisms, as well as at lysine 9 in HeLa cells, although the responsible HMTase(s) have yet to be defined. To

investigate the site utilization profile of *Suv39h1*, unmodified peptides comprising the wild-type H3 N-terminus (aa 1-20) and a mutant K9L peptide, changing lysine 9 to leucine were tested as substrates. Additionally, insulin and peptides comprising the N-termini of CENP-A and macroH2A were included. Peptides were *in vitro* methylated by GST-*Suv39h1*(aa 82-412), and reaction products were separated by high percentage SDS-PAGE and visualized by fluorography. These *in vitro* assays revealed selective methylation of the wild-type H3 peptide, whereas no signals were detected with the CENP-A or macroH2A peptides, or with insulin (*see* FIG. 10B). Importantly, the mutated H3 (K9L) peptide was not a substrate, suggesting that lysine 9 of the H3 N-terminus is a preferred residue for *Suv39h1*-dependent HMTase activity. To more definitively determine this site preference, the wild-type H3 N-terminal peptide was *in vitro* methylated by GST-*Suv39h1*(aa 82-412), using S-adenosyl-(*methyl*-^{3}H)-L-methionine. The labeled peptide, purified by reverse-phase HPLC, was then directly microsequenced, and ^{3}H-incorporation associated with each individual amino acid was analyzed by scintillation counting. The results confirmed selective transfer of methyl-label to lysine 9 (*see* FIG. 10C), demonstrating that *Suv39h1* is a highly site-specific HMTase for the H3 N-terminus *in vitro*.

The invention may be practiced otherwise than as particularly described in the foregoing description and examples.

Numerous modifications and variations of the present invention are possible in light of the above teachings and, therefore, are within the scope of the appended claims.

The entire disclosure of all publications (including patents, patent applications, journal articles, laboratory manuals, books, or other documents) cited herein are hereby incorporated by reference.

References

Aagaard, L., *et al., EMBO J. 18*:1923-1938 (1999)

Aasland, R., and Stewart, A.F., *Nucl. Acids Res. 23*:3168-3174 (1995)

Allshire, R.C., *et al., Genes Dev. 9*:218-233 (1995)

Altschul, S.F., *et al., Nucl. Acids Res. 25*:3389-3402 (1997)

Baksa, K., *et al., Genetics 135*:117-1125 (1993)

Ball, L.J., *et al., EMBO J. 16*:2473-2481 (1997)

Birney, E., *et al., Nucl. Acids Res. 24*:2730-2739 (1996)

Chen, D., *et al., Science 284*:2174-2177 (1999)

Cléard, F., *et al., EMBO J. 16*:5280-5288 (1997)

De Rubertis, F., *et al., Nature 384*:589-591 (1996)

Eddy, S.R., *Genetics 131*:345-352 (1998)

Ekwall, K., *et al., J. Cell. Sci. 109*:2637-2648 (1996)

Frishman, D., and Argos, P., *Proteins, 27*:329-335 (1997)

Grunstein, M., *Cell 93*:325-328 (1998)

Henikoff, S., "Position effect variegation in Drosophila: recent progress," in *Epigenetic mechanisms of gene regulation.* CSHL press (1997)

Ivanova, A.V., *et al., Nat. Genet. 19*:192-195 (1998)

Jacobson, S., and Pillus, L., *Curr. Opin. Genet. Dev. 9*:175-184 (1999)

Jenuwein, T., *et al., Cell. Mol. Life Sci. 54*:80-93 (1998)

Karpen, G.H., and Allshire, R.C., *TIG 13*:489-496 (1997)

Klein, R.R., and Houtz, R.L., *Plant Mol. Biol. 27*:249-261 (1995)

Koonin, E.V., *et al., Nucl. Acids Res. 23*:4229-4233 (1995)

Laible, G., *et al., EMBO J. 16*:3219-3232 (1997)

Larsson, J., *et al., Genetics 143*:887-896 (1996)

Martzen, M.R., *et al, Science 286*:1153-1155 (1999)

Melcher, M., *et al., Mol. Cell Biol. 20*:3728-3741 (2000)

Pehrson, J.R., and Fried, V.A., *Science 257*:1398-1400 (1992)

Platero, J.S., *et al., EMBO J. 14*:3977-3986 (1995)

Reuter, G., and Spierer, P., *BioEssays 14*:605-612 (1992)

Sassone-Corsi, P., *et al., Science 285*:886-891 (1999)

Schotta, G., and Reuter, G., *Mol. Gen. Genet, 262*:916-920 (2000)

Schultz, J., *et al., Nucl. Acids Res. 28*:231-234 (2000)

Strahl, B.D., and Allis, C.D., *Nature 403*:41-45 (2000)

Strahl, B.D., *et al., Proc. Natl. Acad. Sci. USA 96*:14967-14972 (1999)

Sullivan, K.F., *et al., J. Cell Biol. 127*:581-592 (1994)

Tkachuk, D.C., et al., *Cell 71*:691-700 (1992)

Tschiersch, B., *et al., EMBO J. 13*:3822-3831 (1994)

Turner, B.M., *Cell. Mol. Life Sci. 54*:21-31 (1998)

Wallrath, L.L., *Curr. Opin. Genet. Dev. 8*:147-153 (1998)

Wei, Y., *et al., Cell 97*:99-109 (1999)

Zheng, Q., *et al., Protein Expr. Purif. 14*:104-112 (1998)